Computer Science and Application 计算机科学与应用, 2026, 16(2), 1-14
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162034

Hans汉斯

基于语义增强与规则引导的弱监督视频异常检测方法

王津秋渝[1]，宋春林[1]，徐旭辉[2]

[1]同济大学电子与信息工程学院信息与通信工程系，上海
[2]同济大学海洋地质国家重点实验室，上海

收稿日期：2025年12月25日；录用日期：2026年1月22日；发布日期：2026年1月29日

摘　要

视频异常检测(Video Anomaly Detection, VAD)旨在从长时间监控视频中自动识别异常事件，是智能安防与智能交通等场景中的关键技术。受限于异常事件的稀有性与标注成本，现有方法多采用弱监督学习范式，但仍普遍面临异常语义表达不足、跨模态对齐失效以及标签噪声导致训练不稳定等问题。针对上述挑战，本文提出基于语义增强与规则引导的SAGE-VAD (Semantic-Augmented & Guided Enhancement for VAD)框架。设计混合提示集成(Hybrid Prompt Ensemble, HPE)机制，融合人工模板与大模型描述，构建高覆盖度的类别原型。并引入帧级规则分数(Teacher Score)作为先验，通过一致性约束抑制噪声激活并优化关键帧筛选。实验结果表明，本文方法在UCF-Crime和XD-Violence数据集上均取得了显著性能提升。其中，在UCF-Crime数据集上，本文法的视频级AUC达到87.47%，在XD-Violence数据集上，视频级AP提升至85.08%，验证了语义增强与规则引导机制在弱监督异常检测任务中的有效性。

关键词

视频异常检测，弱监督学习，视觉语言预训练模型，多维语义提示

Weakly-Supervised Video Anomaly Detection Method Based on Semantic Augmentation and Rule-Guided Learning

Jinqiuyu Wang[1], Chunlin Song[1], Xuhui Xu[2]

[1]Department of Information and Communication Engineering, College of Electronic and Information Engineering, Tongji University, Shanghai
[2]State Key Laboratory of Marine Geology, Tongji University, Shanghai

Received: December 25, 2025; accepted: January 22, 2026; published: January 29, 2026

文章引用: 王津秋渝, 宋春林, 徐旭辉. 基于语义增强与规则引导的弱监督视频异常检测方法[J]. 计算机科学与应用, 2026, 16(2): 1-14. DOI: 10.12677/csa.2026.162034

Abstract

Video Anomaly Detection (VAD) seeks to automatically detect abnormal events in long-duration surveillance videos and plays a critical role in applications such as intelligent surveillance and smart transportation. Owing to the rarity of anomalous events and the prohibitive cost of fine-grained annotations, most existing methods rely on weakly supervised learning. Nevertheless, they often struggle with limited anomaly semantic expressiveness, suboptimal cross-modal alignment, and unstable optimization induced by noisy supervision. To address these challenges, this paper proposes the Semantic-Augmented & Guided Enhancement for Video Anomaly Detection (SAGE-VAD) framework. First, we design a Hybrid Prompt Ensemble (HPE) mechanism that integrates manual templates with multi-dimensional descriptions generated by LLMs to construct high-coverage category prototypes. And Frame-level Teacher Scores are incorporated as rule-based priors to impose consistency constraints, thereby suppressing noise activations and optimizing keyframe selection in the selector branch. Experimental results demonstrate that SAGE-VAD achieves significant performance gains on the UCF-Crime and XD-Violence datasets, reaching a video-level AUC of 87.47% and an Average Precision of 85.08%, respectively. These results validate the effectiveness of the proposed semantic augmentation and rule-guided mechanisms in weakly-supervised anomaly detection tasks.

Keywords

Video Anomaly Detection, Weakly-Supervised Learning, Cross-Modal Pre-Trained Models, Multi-Dimensional Semantic Prompts

1. 引言

随着信息技术的不断发展，视频数据已成为描述现实场景的重要信息媒介，在公共安全、娱乐内容分析、教育监控以及医疗辅助等领域发挥着关键作用。近年来，数字社会与智慧城市建设的推进使得监控设备在多种复杂环境中得到大规模部署，持续生成数量庞大的监控视频数据，从而对异常事件的自动化检测提出了更高的效率与准确性要求。视频异常检测(Video Abnormal Detection, VAD)旨在自动分析监控视频数据，识别并准确定位可能存在的异常事件[1]。在实际应用中，获取精确的帧级异常标注往往代价高昂且难以规模化，因此近年来弱监督视频异常检测(Weakly Supervised Video Anomaly Detection, WSVAD)受到了广泛关注。WSVAD 方法在训练阶段仅依赖视频级标签或局部标注，相较于全监督方法具有更强的灵活性与可扩展性。Sultani [2]等人率先提出利用视频级弱标签进行异常检测，并构建了包含多种真实犯罪行为的 UCF-Crime 数据集，为 WSVAD 研究奠定了重要基础。随后，Wu [3]等人提出的 XD-Violence 数据集进一步引入音频模态，推动异常检测从单一视觉建模向多模态理解方向发展。

在 WSVAD 中，异常检测器需要在仅提供视频级标注的情况下产生帧级异常置信度，该领域目前的大多数研究遵循以下系统化流程，其中第一步是使用预训练的视觉模型如 C3D [4]、I3D [5]及 ViT [6]等提取帧级特征，然后将这些特征输入到基于多示例学习(MIL)的二分类器中进行模型训练，最后一步是根据预测的异常置信度检测异常事件。尽管这种基于分类的方案实现简单且效果较好，但其并未充分地利用跨模态的关系，例如视觉－语言关联，视觉－音频关联等。近年来，视觉－语言预训练模型的快速发展为 VAD 任务提供了新的研究视角。以 CLIP [7]为代表的模型通过在大规模图文对上进行对齐学习，获

得了具有良好泛化能力的跨模态语义表示，并在多种视觉理解任务中取得了显著性能提升。得益于OpenAI 在海量噪声图文数据上的预训练，这类模型展现出强大的语义建模能力，为异常行为的高层语义理解提供了重要先验[8]。

直接将以图像为中心设计的 CLIP 模型应用于 WSVAD 任务，仍面临文本语义覆盖受限、跨模态融合不足以及规则知识难以有效利用等挑战。为此，本文提出一种融合语义增强与规则引导的弱监督视频异常检测框架，通过显式建模异常行为的多维语义结构并引入训练与推理阶段的规则一致性约束，在仅依赖视频级监督的条件下提升检测的稳定性与鲁棒性。

2. 基于语义增强与规则引导的视频异常检测框架

受到近年来，VLM 适配下游任务的 Prompt-Learning 研究启发[9]，本文将注意力转向更深层次的文本侧增强与跨模态一致性建模，并提出了一个全新的框架 SAGE-VAD (Semantic-Augmented & Guided Enhancement for VAD)，图 1 为 SAGE-VAD 的网络结构框架示意图。

针对现有方法中普遍采用单一或少量人工模板、难以覆盖异常行为多维语义结构的问题[10]，本文从异常语义本身的结构性问题出发，对文本提示建模方式进行了系统性的增强。不同于将异常类别视为单一语义标签[11]，本文认为异常行为本质上由于多个语义维度共同构成，包括行为类型、突发变化、正常性偏差、影响程度以及人体动作与场景属性。因此基于这个认知，本文提出了一种多维异常语义提示构建算法(Hybrid Prompt Ensemble, HPE)，将任务相关的人工模板与大模型生成的多维语义相组合，从而在本文侧显式构建异常行为的高维语义空间。通过对多源 Prompt 编码结果进行聚合，模型不再依赖某一特定模板的偶然表达，而是学习到更加稳定且具有泛化性的类别语义原型，为后续的跨模态对齐提供了坚实的语义基础。

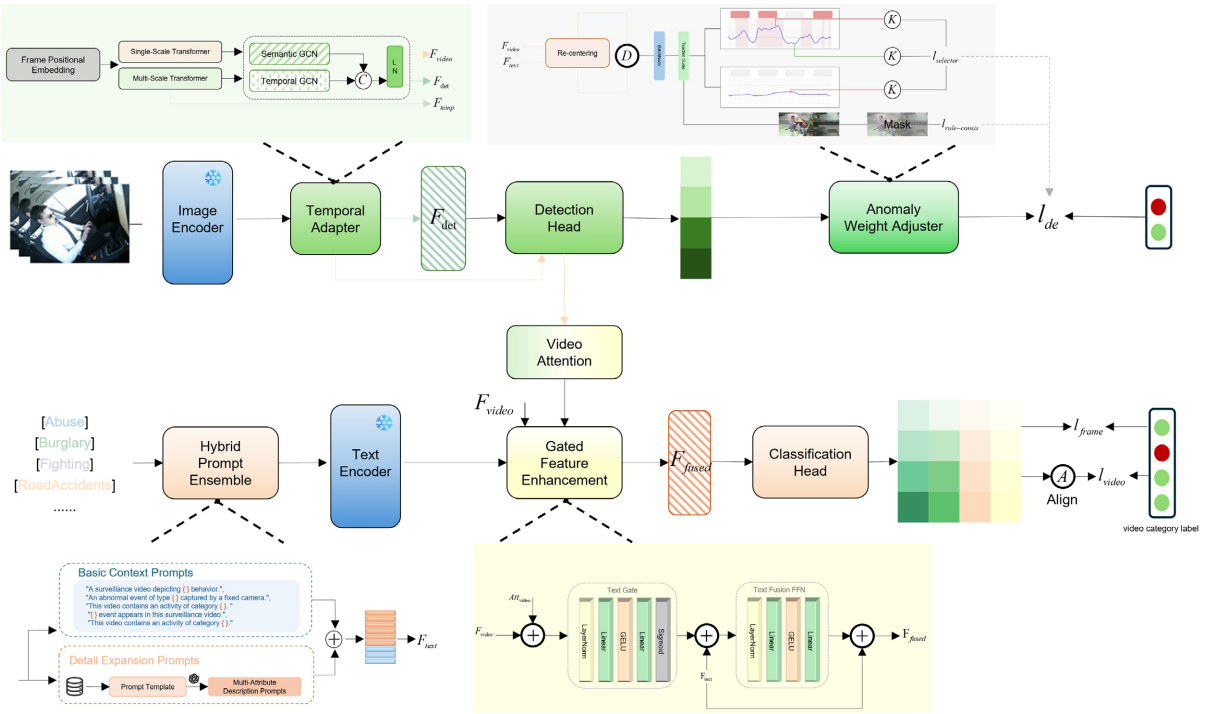

Figure 1. Overall architecture of the SAGE-VAD framework
图 1. SAGE-VAD 网络结构框架示意图

其次，针对弱监督异常行为检测设定下的标签噪声大、训练过程不稳定等问题，本文通过基于规则

引导的异常帧选择算法将规则引入到 WSVAD 的训练和推理全过程中，不同于仅在后处理阶段使用规则过滤预测结果，本文将规则与语言模型联合推理得到的帧级异常评分作为一种软监督信号，并在高置信区域对模型输出施加一致性约束。通过设计渐进式引入的规则一致性损失，模型在保持数据驱动学习能力的同时，能够在明确的正常或者异常区域受到规则知识的引导，从而有效地抑制噪声样本带来的干扰。此外，规则信息还被用于引导关键帧的选择与推理阶段的预测修正，让模型在弱监督与开放场景下获得更加稳健的异常判别能力。

综上，本文通过在文本语义建模、跨模态融合机制以及规则知识引导三个层面进行协同设计，系统性地缓解了 CLIP 在弱监督场景下面临的语义不足、对齐失效与鲁棒性较差等问题，显著提升了模型在复杂监控场景下的异常识别性能与语义泛化能力。

3. 基于多维异常语义提示的语义增强算法

本文将以 CLIP 作为研究的基础 I-VLMs。在预训练阶段，每个批次由图像－文本对组成，图像编码器和文本编码分别为图像和文本计算特征嵌入表示，随后对所有的图像－文本对计算其余弦相似度矩阵。训练目标是通过最大化正确匹配对之间的相似度、最小化错误匹配对之间的相似度，从而联合优化图像编码器和文本编码器。在推理阶段，CLIP 根据图像特征 $f(x)$ 和文本特征 $g(t)$ 之间的相似度，以零样本的方式对文本进行分类。给定一幅图像 x 以及一组类别文本描述 $\{y_c\}_{c=1}^{C}$，图像编码器 $f_{img}(\cdot)$ 与文本编码器 $f_{text}(\cdot)$ 分别将其映射到共享的语义嵌入空间

$$v = f_{img}(x),\ t_c = f_{text}(y_c) \tag{1}$$

其中 $v, t_c \in \mathbb{R}^D$，CLIP 基于图像特征与文本特征之间的余弦相似度进行预测，类别得分/预测可表示为

$$s_c = \cos(v, t_c) \tag{2}$$

$$\hat{c} = \arg\max_c s_c \tag{3}$$

尽管公式(2)够清晰地刻画出 CLIP 的零样本推理流程，但其有效性隐含了一个关键的假设：类别文本特征能够充分、稳定地表征对应类别的语义空间。然而，在视频异常检测任务中，该假设往往难以成立。异常行为通常具有多维语义属性，例如行为类型、突发性、正常性偏离程度以及潜在影响等，仅依赖单一人工设计的 Prompt 模板，难以全面地覆盖这些语义维度。同时，已有研究表明，I-VLM 对 Prompt 的措辞形式具有高敏感性，其也进一步地削弱了对应模型在固定领域异常场景下的泛化能力。

为缓解上述问题，本文提出 Hybrid Prompt Ensemble (HPE)，如图 2，通过结合人工设计模板与大语言模型生成的多样化语义描述，构建更鲁棒的类别文本。具体而言，对于每个类别将不再使用单一文本描述，而是构建一个 Prompt 集合，该集合同时包含人工模板 Prompt 与 LLM 生成的语义 Prompt，用于覆盖异常行为的多维语义表达：

$$P_c = \{p_{c,1}, p_{c,2}, \cdots, p_{c,N_c}\} \tag{4}$$

其中 $P_{c,i}$ 是为类别 c 设计的第 i 个 Prompt，同时包含人工模板 Prompt 与 LLM 生成的语义 Prompt，用于覆盖异常行为的多维语义表达。利用 CLIP 文本编码器 $f_{text}(\cdot)$ 每个 Prompt 被映射为文本嵌入：

$$f_{text}(p_{c,i}),\ i = 1, \cdots, N_c \tag{5}$$

随后通过对同一类别下的多 Prompt 表征进行聚合，得到按照均值融合的类别级文本特征见公式(6)，

$$t_c = \frac{1}{N}\sum_{i=1}^{N_c} f_{c,i} \tag{6}$$

在引入 Hybrid Prompt Ensemble 后 CLIP 的预测形式仍保持与公式(2)保持一致，但其中文本特征由公式(6)所定义的多 Prompt 融合原型所替代。最终，模型的类别预测可表示为公式(7)

$$s_c = \cos\left(v, \frac{1}{N_c}\sum_{i=1}^{N_c} f_{text}\left(p_c, i\right)\right) \tag{7}$$

通过这种方式，模型将不再依赖某一特定 Prompt 的偶然表达，而是学习到更稳定、具有更强语义覆盖能力的类别类型，为后续跨模态对齐与异常判别提供了更加可靠的语义基础。

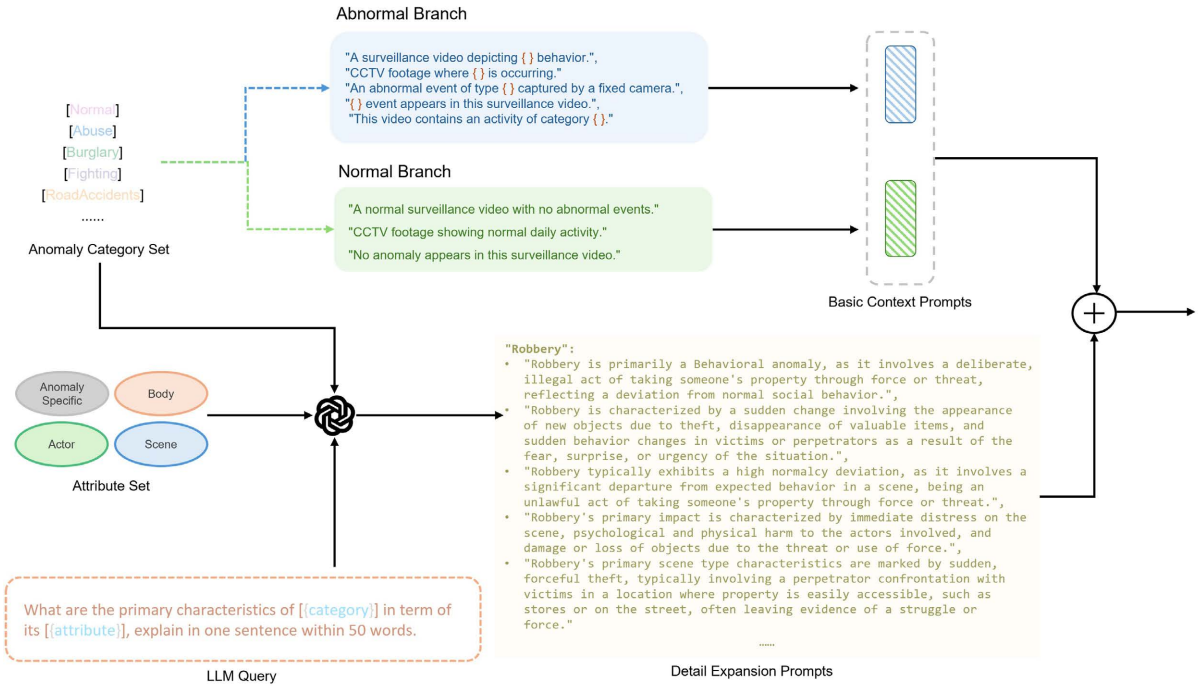

Figure 2. Text feature enhancement algorithm based on HPE
图 2. 基于混合提示机制的文本特征增强算法

4. 基于规则引导的异常帧选择算法

在视频异常检测任务中，如何有效地选择与异常判断相关的关键帧，避免噪声标签对模型的影响，以及全面、准确地理解异常行为的多维特征，长期以来都是研究中的难点。VADCLIP 中使用 Top-K 选择基于二分类异常置信度来确认哪些视频中最有可能是异常，选择出 K 个最有可能异常的帧，如公式(8)，其中 p 为帧的置信度分数。然后进行二元交叉熵损失，如公式(9)，其主要目的是通过选取置信度最高的帧，粗粒度的检测出视频中的异常。

$$A = \text{Top-K}\left(p\right) \tag{8}$$

$$\mathcal{L}_{bce} = -\sum\left(y\cdot\log\left(A\right)+\left(1-y\right)\cdot\log\left(1-A\right)\right) \tag{9}$$

这种方法基于全局特征对齐，虽然可以提高检测精度，但是会存在由于无法精确的控制与异常判断无关的帧数，以及没有进一步的优化异常帧的选择，导致模型在面对复杂数据时容易受到干扰等问题。

4.1. 基于规则的帧级异常可信度增强

本文中新提出的 Anomaly Weight Adjuster 首先引入了 Teacher Score 引导，通过对每一帧的 Teacher

Score 进行调整，来加权和选择最相关的帧。Teacher Score 由规则系统生成，如图 3 所示。具体而言，规则系统在帧级推理阶段使用的 VLM 模型为 Mistral-7B-Instruct-v0.2，该模型采用 Transformer 架构，具备强大的推理和规则遵循能力，更适合在大基准数据集上进行大规模且经济的推理测试。规则系统在规则生成阶段，由于规则生成对于逻辑的要求相对较高且调用的频率较低，因此选择调用 GPT 4 的 API 完成规则生成任务。

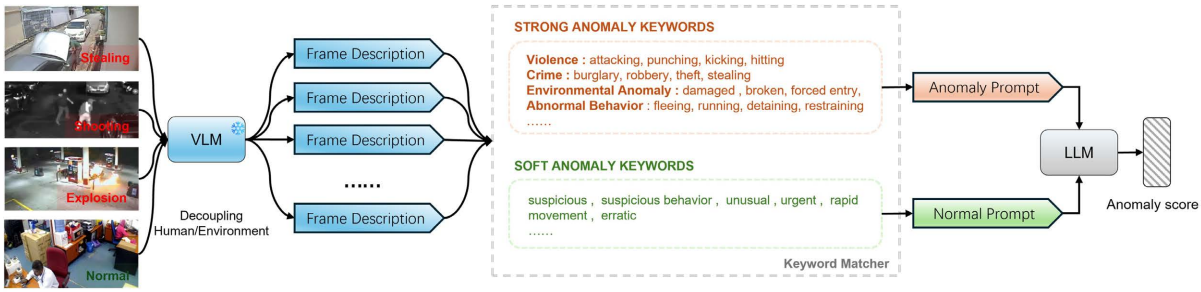

Figure 3. Overview of the Teacher Guide workflow
图 3. Teacher Guide 流程示意图

该系统基于 VLM 模型产生的帧级描述后，首先进行异常关键词库的快速过滤匹配。在此匹配前，关键词将被划分为强关键词与弱关键词两类，前者将指向明确的犯罪/暴力等危险行为，如盗窃、攻击、破坏等，用于匹配后触发更严格的异常行为复核。而后者多为异常弱线索，如可疑、快速移动等，仅作为提示信息，不直接构成异常判定依据。基于此过滤分流策略，系统实现了双重验证逻辑，亦见图 3。第一步用关键词精准匹配实现高效的召回与分流，第二步在注入提前生成的规则文本的约束下调用对应的 Prompt 文件进行复核，从而保证在帧级可扩展的同时，显著抑制由描述噪声或者帧级噪声带来的误判。

最后，对每一帧信息 t，规则系统会输出一个连续的 Teacher Score $r_t \in [0,1]$，该分数衡量当前帧发生异常行为的可信度，其中较高的 r_t 表示该帧在规则层面更可能与异常事件相关。与直接使用规则结果进行硬判决不同，Teacher Score 仅作为软可信信号，用于引导模型的帧级语义选择过程，详见图 4。

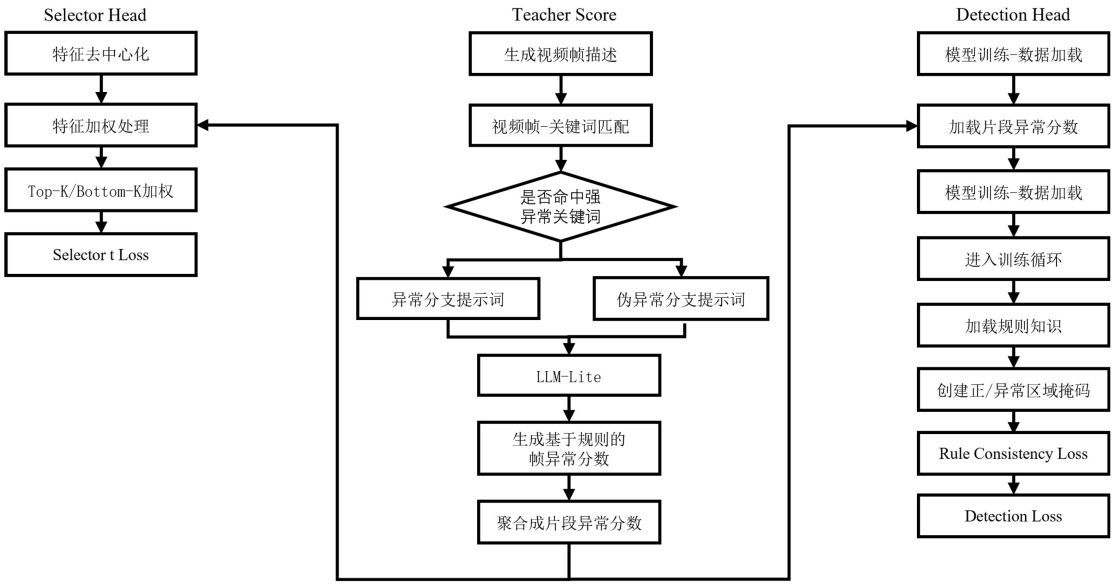

Figure 4. Dual-path training guidance via rule-based frame-level anomaly confidence modeling
图 4. 基于规则的帧级异常可信度建模及双路径训练引导流程图

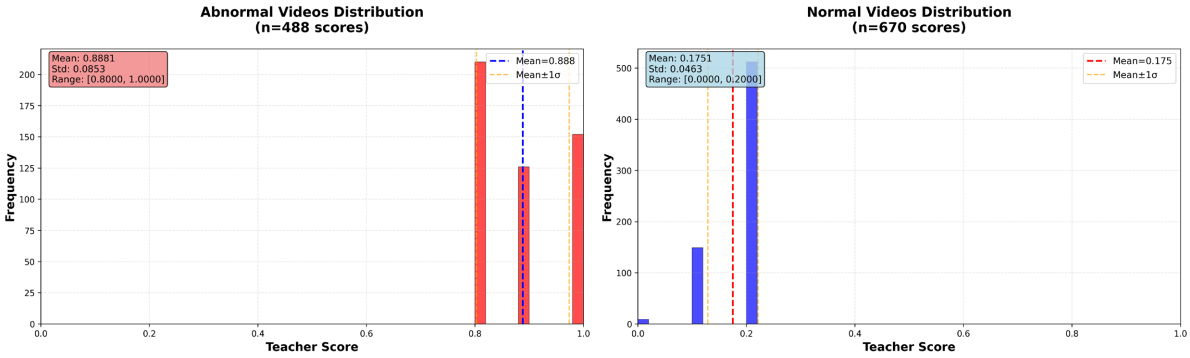

Figure 5. Distribution of frame-level Teacher Score
图 5. 帧级 Teacher Score 分布示意图

本文对正常视频与异常视频中生成的 Teacher Score 进行了分析。如图 5 所示，在正常视频中，大多数帧的 Teacher Score 集中分布于较低区间，表明规则系统倾向于对正常行为给出较低的异常可信度；而在异常视频中，Teacher Score 则显著集中于高分区间，呈现出良好的分布可分性。这一现象表明，基于规则与语义提示构建的分数能够在统计层面有效刻画帧级异常可信度，为模型提供稳定且可靠的先验信息。

4.2. 规则一致性引导的帧级检测约束

在弱监督视频异常检测任务中，仅依赖视频级标签对检测分支进行训练，容易导致模型在大量正常背景帧上产生异常误激活，从而削弱帧预测的稳定性。为缓解这一问题，本节进一步将 Teacher Score 作为帧级异常可信度先验，引入到检测分支的学习过程中，对帧级异常预测施加规则一致性约束。具体而言，假设检测分支在第 b 个视频，第 t 个片段上的异常预测概率为公式(10)

$$p_{\text{det}}^{(b,t)} = \sigma\left(\text{logits}_{\text{det}}^{(b,t)}\right) \tag{10}$$

其中 $\sigma(\cdot)$ 表示 sigmoid 函数，其对应的 Teacher Score 记为 $S_{teacher}^{(b,t)} \in [0,1]$，用于衡量规则系统对该帧发生异常行为的可信度。

正常区域建模：考虑到规则系统在异常区域可能存在噪声或不完整性，本文仅在 Teacher Score 高确信其为正常的帧上引入一致性约束，具体地，定义正常区域掩码为

$$m_{normal}^{(b,t)} = \begin{cases} 1, & s_{teacher}^{(b,t)} \le \tau_{normal} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

其中 τ_{normal} 为正常区域阈值，在实验中取得是 $\tau_{normal} = 0.2$，该掩码刻画了规则系统在帧/片段层面高度确认为正常的区域，这些区域被视为可靠的负约束位置。

在上述高置信正常区域内，本文引入了规则一致性损失(Rule Consistency Loss)，以抑制检测分支在这些区域产生异常响应，见公式(12)

$$\mathcal{L}_{rule-nor} = \frac{\sum_{b=1}^{B}\sum_{t=1}^{T} p_{\text{det}}^{(b,t)} \cdot \left(m_{normal}^{(b,t)}\right)^2}{\sum_{b=1}^{B}\sum_{t=1}^{T} m_{normal}^{(b,t)} + \varepsilon} \tag{12}$$

其中 $\varepsilon = 10^{-6}$ 用于数值稳定性。上述损失函数等价于对规则高度确信为常的帧，最小化异常预测概率的平方期望，异常区域的约束将在后续的工作中进一步的讨论。目前是通过 $\lambda_{abn} = 0$ 进行了约束，综合这两项，规则一致性损失的定义为

$$\mathcal{L}_{rule} = \alpha_{scale}\left(\lambda_{nor}\mathcal{L}_{rule-nor} + \lambda_{abn}\mathcal{L}_{rule-abn}\right) \tag{13}$$

4.3. 帧级可信度与关键帧选择优化算法

本文将 Teacher Score 作为帧级可信度先验引入语义建模过程，通过对帧–类别语义 logits 进行自适应重加权，实现对“语义相关性”与“样本可信度”的联合建模。在此基础上，进一步执行差异化的关键帧选择与帧级优化策略。对于视频 b 的片段 s，Teacher Score $r_{b,s}$ 位于 $[0,1]$ 区间。对于异常视频($b < B/2$)

$$w_{b,s,c} = \begin{cases} 1+\alpha\cdot\left(1-r_{b,s}\right), & c=0\ (\text{正常类}) \\ 1+\alpha\cdot r_{b,s}, & c\geq 1\ (\text{异常类}) \end{cases} \tag{14}$$

对于正常视频($b < B/2$)

$$w_{b,s,c} = \begin{cases} 1+\alpha\cdot r_{b,s}, & c=0\ (\text{正常类}) \\ 1+\alpha\cdot\left(1-r_{b,s}\right), & c\geq 1\ (\text{异常类}) \end{cases} \tag{15}$$

调整后的 logits 为公式(16)

$$\tilde{S}_{b,s,c} = w_{b,s,c}\cdot S_{b,s,c} \tag{16}$$

Anomaly Weight Adjuster 通过引入 Teacher Scores 引导选择和自适应调整机制，增强模型对异常帧的精准选择能力。同时模型可根据 Teacher Scores 动态调整每个视频帧的权重，聚焦关键帧，有效避免无关帧的干扰。

4.4. 多维度关键帧选择策略

与其他研究仅依赖异常帧的 Top-K 选择不同，Anomaly Weight Adjuster 提出了更精细的 Top-K/Bottom-K 选择，具体流程见图 6。其核心是在帧级语义空间中显式建模异常相关性与样本可信度，并据此进行分层筛选。具体而言，给定视频 b 的帧级视觉特征 $X_{b,t}$ 与类别文本特征 U_c，首先通过去中心化与归一化操作，计算帧–类别之间的语义相似度。此相似度刻画了每一帧在语义层面上与各异常类别的匹配程度：

$$S_{b,t,c} = \left\langle \hat{X}_{b,t}, \hat{U}_c \right\rangle \tag{17}$$

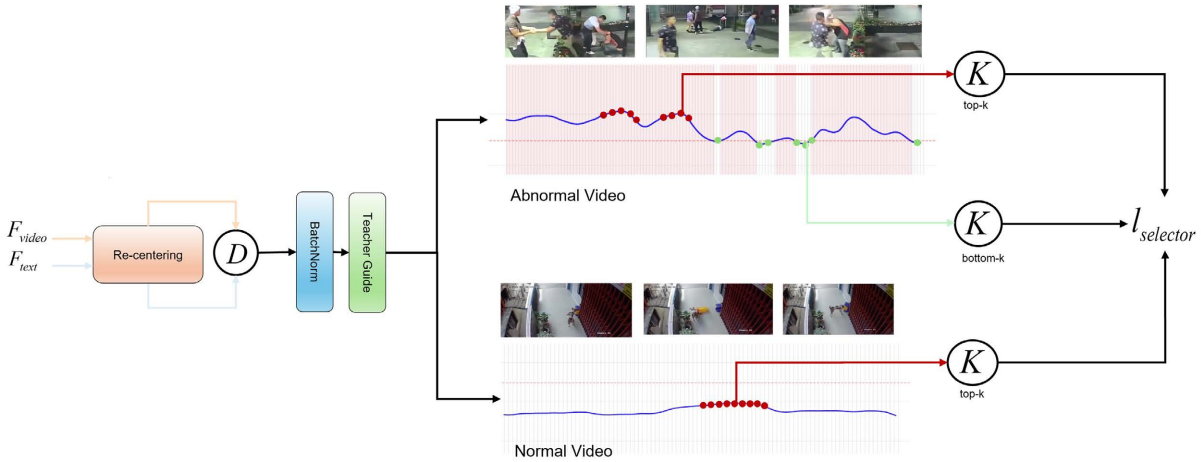

Figure 6. Multi-dimensional keyframe selection mechanism
图 6. 多维度关键帧选择示意图

与其他研究仅依赖异常帧的 Top-K 选择不同，Anomaly Weight Adjuster 提出了更精细的 Top-K/Bottom-K 选择策略，通过多维度异常行为提示与多种选择策略提高模型的性能。其核心是在帧级语义空间中显式建模异常相关性与样本可信度，并根据此进行分层筛选。具体而言，给定视频 b 的帧级视觉特征 $X_{b,t}$ 与类别文本特征 U_c，首先通过去中心化与归一化操作，计算帧－类别之间的语义相似度。此相似度刻画了每一帧在语义层面上与各异常类别的匹配程度：

$$S_{b,t,c} = \left\langle \hat{X}_{b,t}, \hat{U}_c \right\rangle \tag{18}$$

在此基础上，引入由规则系统提供的 Teacher Score $r_{b,t} \in [0,1]$，以衡量每一帧发生异常的可信度。根据视频标签(正常/异常)，对语义 logits 进行自适应重加权，见公式，*其中* $w_{b,t,c}$ 通过 $r_{b,t}$ 调节，使得高置信异常帧在异常类别方向被增强，而低置信帧被抑制。该步骤实现了从纯相似度排序到语义与可信度联合建模的转变。随后，基于调整后的语义得分 $\tilde{S}$，在不同视频类型下执行差异化的选择策略。对于异常视频，通过 Top-K 选择得到最具异常判断力的片段集合

$$\mathcal{I}_b^{topk-abn} = TopK\left(\tilde{S}_{b,:,c_b}, K_{top}\right) \tag{19}$$

同时引入 Bottom-K 选择，找出视频中可能存在的噪声或弱异常片段

$$\mathcal{I}_b^{bottomk-abn} = BottomK\left(\tilde{S}_{b,:,c_b}, K_{bottom}\right) \tag{20}$$

对于正常视频 $b \in [B/2, B]$，选择异常类别维度上通过 Top-K，选取最具“伪异常倾向”的片段作为负样本，用于对比学习

$$\mathcal{I}_b^{topk-nor} = TopK\left(\sum_{c>1} \tilde{S}_{b,:,c_b}, K_{top}\right) \tag{21}$$

通过上述设计，Anomaly Weight Adjuster 不再仅关注最异常的少数帧，而是同时学习异常核心片段、异常边界片段以及正常视频中的困难负样本，从而在帧级层面构建更稳定的判别边界，最终这些被选中的片段被送进选择损失函数进行联合优化，使得梯度主要作用于语义最相关、置信度最高且信息互补的关键帧集合。

Selector Loss 由三个互补的子损失项构成，均基于负对数似然损失进行计算，包括异常视频 Top-K 片段损失、异常视频 Bottom-K 片段损失以及正常视频 Top-K 片段损失分别对应公式(22)~(24)

$$\mathcal{L}_{topk-abn} = -\frac{1}{N}\sum_{i=1}^{N} \log P\left(y_i \mid x_i\right) \tag{22}$$

其中 $N = B_{abn} \times num_{topk} \times seg\_length$， y_i 为该异常视频对应的真实异常类别标签，该损失用于强化模型对异常核心片段的语义判别能力。

$$\mathcal{L}_{bottomk-abn} = -\frac{1}{M}\sum_{j=1}^{M} \log P\left(normal \mid x_j\right) \tag{23}$$

其中 $M = B_{abn} \times num_{bottomk} \times seg\_length$，所有 Bottom-K 项均被视为正常类，该项损失显式约束异常视频中语义响应较低的片段，抑制异常类别在噪声区域的误激活，从而提升模型的边界稳定性。

$$\mathcal{L}_{topk-nor} = -\frac{1}{K}\sum_{k=1}^{K} \log P\left(normal \mid x_k\right) \tag{24}$$

其中 $K = B_{nor} \times num_{topk} \times seg\_length$，该损失项促使模型在面对看起来像正常的异常片段的时候，仍能够保持正常的判断，从而显著的降低误报率。

通过公式(25)的 Selector Loss 设计，模型在训练过程中被显式引导关注三类关键片段：异常核心片段、异常边界片段以及正常视频中的困难负样本片段。这种分层约束方式使得梯度主要集中于语义最具判别力的位置，从而在弱监督场景下显著提升异常检测的稳定性与泛化能力。

$$\mathcal{L}_{selector} = w\left(\lambda_{ta}\mathcal{L}_{topk-abn} + \lambda_{ba}\mathcal{L}_{bottomk-abn} + \lambda_{tn}\mathcal{L}_{topk-nor}\right) \tag{25}$$

5. 实验与分析

5.1. 实验设置与关键指标分析

用于验证本文方法有效性的实验均在公开视频异常检测数据集上进行，主要包括 UCF-Crime 与 XD-Violence 两个具有代表性的真实世界监控视频数据集。所有实验均在统一的训练配置与评测流程下完成，以确保结果的公平性与可复现性。

在模型训练阶段，本文采用 CLIP 作为基础视觉–语言特征提取器，CLIP 主干参数保持冻结，仅对新增的语义增强模块、检测分支及选择模块进行训练。模型训练共进行 10 个 epoch，以避免在弱监督条件下发生过拟合。优化策略选用 AdamW，初始学习率设置为 2×10^{-5}，权重衰减采用默认配置 1×10^{-4}，并结合学习率调整调度策略以提升训练的稳定性。为保证实验结果的可复现性，所有实验均固定随机种子 3407 进行。

在批次设置方面，对 UCF-Crime 和 XD-Violence，训练批次大小分别设置为 64 和 96。为了更好地保证该方法在不同随机初始化条件下的稳定表现，实验基于 5 个不同随机种子 3407、2026、999、42、7 进行独立训练并记录均值 ± 标准差。在损失函数设计上，除视频级主损失外，本文引入帧/片段级多实例学习辅助损失，其权重为 0.05 用户在训练过程中提供额外的片段、帧级别监督信号。对于规则一致性约束与关键帧选择相关损失，采用渐进式 warm-up 策略引入，以降低规则先验在训练初期对模型优化的干扰。

在训练环境方面，本文基于 PyTorch 深度学习框架及进行实现，CUDA 版本为 12.8，cuDNN 版本为 9.0.1.2，所有实验均在 NVIDIA GeForce RTX 5090 GPU 上完成，该设备配备 32 GB 显存，能够满足长时序视频特征建模与多分支训练的显存需求。

Table 1. Comparison of video anomaly detection performance under different supervision settings
表 1. 不同监督范式下视频异常检测方法的性能对比

Supervision	Method	Feature	UCF-Crime AUC (%)	XD-Violence AP
Semi-Supervised	GODS [12]	I3D	70.46	-
	GCL	ResNext	74.20	-
Zero Shot	CLIP-TSA [13]	CLIP	87.58	82.17
Weakly-Supervised	Ju *et al.*	-	84.72	76.57
	Sultani *et al.*	-	84.41	75.81
	RTFM [14]	I3D	84.30	78.27
	AVVD [15]	CLIP	82.45	78.10
	DMU [16]	CLIP	86.75	82.41
	UMIL [17]	X-CLIP	86.75	-
	PLOVAD [18]	CLIP	87.06	-
	SAGE-VAD (Ours)	CLIP	87.75 (±0.04)	85.08 (±0.11)

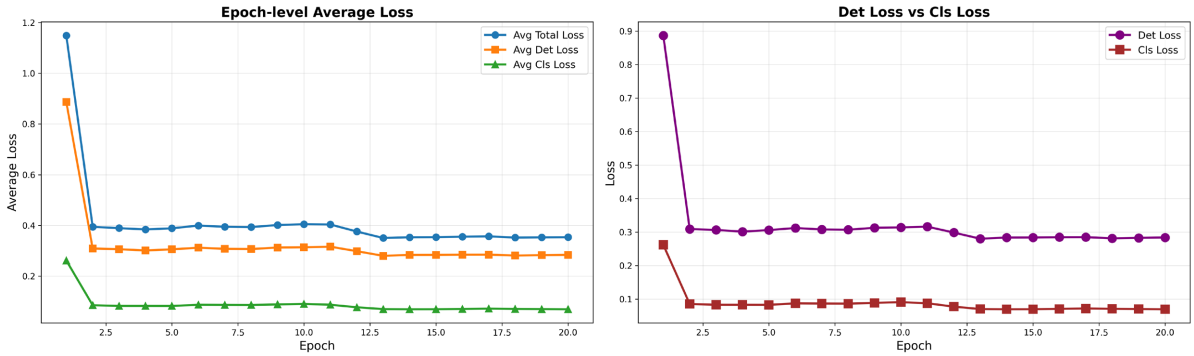

Figure 7. Training Loss curves on the XD-Violence dataset
图 7. 在 XD-Violence 数据集上的训练 Loss 曲线

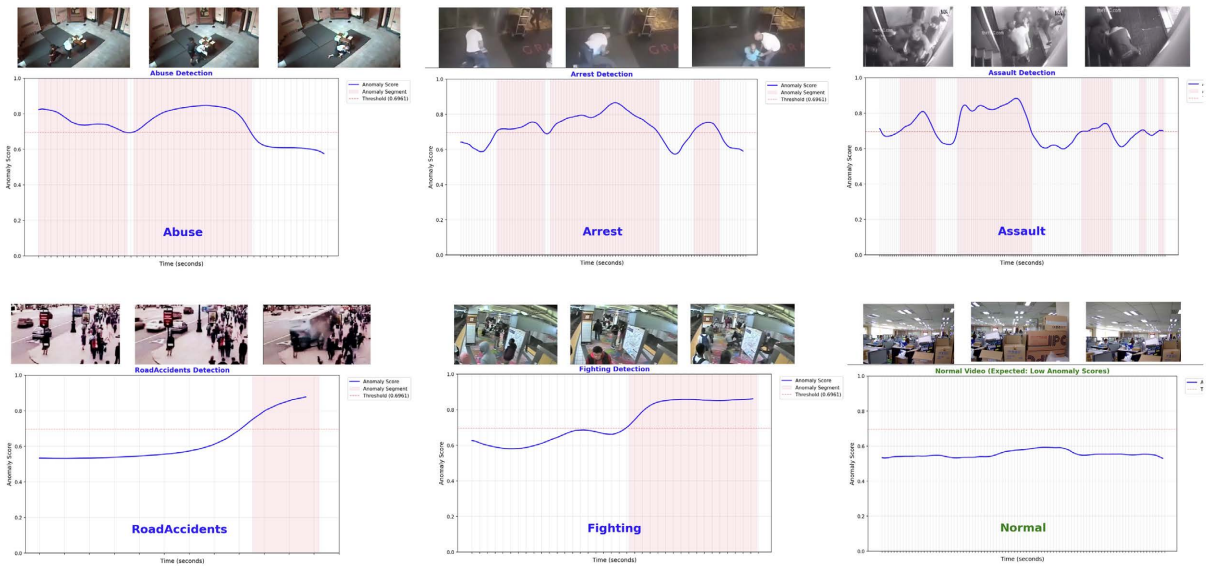

Figure 8. Visualization of frame-level anomaly responses over time for different anomaly categories
图 8. 不同异常类别的视频帧级异常响应随时间变化的可视化结果

表 1 展示了 SAGE-VAD 与现有代表性方法在 UCF-Crime 和 XD-Violence 数据集上的性能对比，图 7，图 8 分别是模型在训练以及测试过程中的可视化结果。通过表 1 可以发现传统的半监督方法(如 GODS、GCL)由于依赖额外标注或特征，其整体性能明显受限。零样本方法在一定程度上受益于视觉语言模型的语义泛化能力，但在复杂场景下仍存在性能瓶颈。在弱监督的设置下，基于 CLIP 的方法整体优于传统 3D CNN 特征方法，验证了跨模态语义建模在视频异常检测中的有效性。在此基础上，SAGE-VAD 在 UCF-Crime 上取得了 87.75% ± 0.04%的 AUC，在 XD-Violence 上达到了 85.08% ± 0.11%的 AP，相比于大多数弱监督方法有较为客观的增益，展示出语义增强与规则引导的优势。

5.2. Hybrid Prompt Ensemble 的有效性分析

从表 2 中可以看到，仅使用人工模板时，模型在 AUC 上能够取得 86.07%，但在 mAP@IoU 指标上整体表现较弱，平均值仅为 4.38%。这表明人工模板虽然能够提供稳定、结构化的类别语义，但是特征表达方式比较单一，难以覆盖异常行为在时序演化和细粒度表现上的多样性，因而对异常片段的精确定位能力有限。而相比仅使用人工模板，仅采用 LLM 生成的 Prompt 描述后，模型在各 IoU 阈值下的 mAP 均有明显提升，平均 mAP 提升至 6.85%，AUC 也提升至 86.92%。该结果表明，LLM 生成的多样化语义描

述能够有效扩展异常语义空间，从而提升模型对异常片段的排序与区分能力。

Table 2. Performance of different Prompt construction strategies based on the UCF-Crime dataset
表 2. 基于 UCF-Crime 数据集的不同 Prompt 构建策略的性能表现

实验名称	mAP@IoU (%)						AUC (%)
	0.1	0.2	0.3	0.4	0.5	AVG	
仅人工模板	8.09	5.88	4.63	1.91	1.23	4.348	86.07
仅 LLM 生成	12.1	9.19	6.77	4.53	1.64	6.846	86.92
人工模板 + LLM 生成	14.45	10.36	7.59	4.45	2.20	7.81	87.75

Table 3. Ablation experiment results of key modules
表 3. 关键模块消融实验结果

HPE	Gated	Anomaly Weight	Teacher Score	AUC (%)	AP (%)
-	-	-	-	85.94	27.02
√	-	-	-	87.02	33.49
√	√	-	-	87.02	36.30
√	√	√	-	87.44	36.72
√	√	√	√	87.75	39.61

最后，将人工模板与 LLM 生成描述相结合构建 HPE 后，模型在所有指标上均取得最优性能，平均 mAP@IoU 提升至 7.81%，AUC 达到 87.75%。这一结果验证了结构化语义约束与多维语义扩展之间的互补性，不仅增强了异常语义原型的表达能力，也提升了跨模态对齐的稳定性，为后续基于语义选择与规则引导的关键帧建模提供了更加可靠的文本特征。

5.3. 关键模块消融与规则引导机制分析

为了系统性地验证本文所提出的所有关键模块的有效性，本文在 UCF-Crime 数据集上展开了消融实验，实验结果如表 3 所示。

表 3 给出了在逐步引入不同模块配置下模型性能的变化情况，其中包括 Hybrid Prompt Ensemble、Anomaly Weight Adjuster 以及 Teacher Score 引导机制。首先在未引入任何增强模块的设置上，模型依赖基于 CLIP 特征提取的弱监督视频级标签进行训练其 AUC 为 85.94%，而 AP 仅为 27.02%。这一结果表明，传统 MIL 框架在帧级异常定位于排序方面的优势有限，容易受到噪声标签的影响。

在逐步引入各关键模块的过程中，模型性能呈现出稳定且一致的提升趋势。首先 HPE 模块显著提升了异常语义建模能力，使模型在保持较高 AUC 的同时，AP 也获得明显提升，验证了多源语义增强对跨模态对齐的有效性。随后，引入 Gated (视频–文本特征基于门控的融合)后，模型的 AP 得到进一步提升，但 AUC 基本保持不变，表明该模块主要改善了帧级异常判别的区分性，而非视频级分类能力。进一步结合 Anomaly Weight Adjuster 后，模型能够更加聚焦于判别性关键帧，从而提升弱监督训练的稳定性。最终，在引入 Teacher Score 作为帧级异常可信度先验后，模型在 AUC 与 AP 上均取得最优结果，验证了规则引导机制在抑制噪声激活、增强异常定位精度方面的有效性。

6. 结论

综合上述实验结果可以看出，SAGE-VAD 在整体性能、文本语义建模、关键帧选择以及规则引导稳

定性等多个维度均展现出一致优势。大量实验结果也可以进一步表明 SAGE-VAD 在 AUC、AP 及 mAP@IoU 等指标上皆取得了较大幅度的增长，验证了本文方法在复杂监控场景下的有效性与泛化能力。模型在 UCF-Crime 与 XD-Violence 等公开基准数据集上分别取得了 AUC 为 87.75%以及 AP 为 85.08%，相比 Baseline 模型性能有了显著的优化。后续的消融实验分析可以进一步表明，各模块并非简单叠加，而是在语义建模、跨模态对齐与弱监督约束层面形成了协同增强关系，从而有效提升了模型在复杂监控场景下的异常检测与定位能力。

基金项目

本文受上海市科委 2024 创新行动计划项目资助，项目编号：24511103302。

参考文献

[1] Liu, Y., Yang, D., Wang, Y., Liu, J., Liu, J., Boukerche, A., *et al*. (2024) Generalized Video Anomaly Event Detection: Systematic Taxonomy and Comparison of Deep Models. *ACM Computing Surveys*, **56**, 1-38. https://doi.org/10.1145/3645101

[2] Sultani, W., Chen, C. and Shah, M. (2018) Real-World Anomaly Detection in Surveillance Videos. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6479-6488. https://doi.org/10.1109/cvpr.2018.00678

[3] Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., *et al*. (2020) Not Only Look, but Also Listen: Learning Multimodal Violence Detection under Weak Supervision. In: *Lecture Notes in Computer Science*, Springer, 322-339. https://doi.org/10.1007/978-3-030-58577-8_20

[4] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/iccv.2015.510

[5] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6299-6308. https://doi.org/10.1109/cvpr.2017.502

[6] Dosovitskiy, A. (2020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/pdf/2010.11929/1000

[7] Radford, A., Kim, J.W., Hallacy, C., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, Online, 18-24 July 2021, 8748-8763.

[8] 张琳, 陈兆波, 马晓轩, 等. 无监督和弱监督视频异常检测方法回顾与前瞻[J]. 科学技术与工程, 2024, 24(19): 7941-7955.

[9] Giambastiani, B.M.S. (2007) Evoluzione Idrologica ed Idrogeologica della Pineta di San Vitale (Ravenna). Ph.D. Thesis, Bologna University.

[10] 苏文浩. 基于弱监督学习的视频异常检测方法研究[D]: [硕士学位论文]. 济南: 山东大学, 2024.

[11] Yao, H., Zhang, R. and Xu, C. (2023) Visual-Language Prompt Tuning with Knowledge-Guided Context Optimization. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 6757-6767. https://doi.org/10.1109/cvpr52729.2023.00653

[12] Wang, J. and Cherian, A. (2019) GODS: Generalized One-Class Discriminative Subspaces for Anomaly Detection. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 8200-8210. https://doi.org/10.1109/iccv.2019.00829

[13] Joo, H.K., Vo, K., Yamazaki, K. and Le, N. (2023) CLIP-TSA: Clip-Assisted Temporal Self-Attention for Weakly-Supervised Video Anomaly Detection. 2023 *IEEE International Conference on Image Processing* (*ICIP*), Kuala, 8-11 October 2023, 3230-3234. https://doi.org/10.1109/icip49359.2023.10222289

[14] Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W. and Carneiro, G. (2021) Weakly-Supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 4955-4966. https://doi.org/10.1109/iccv48922.2021.00493

[15] Wu, P., Liu, X. and Liu, J. (2023) Weakly Supervised Audio-Visual Violence Detection. *IEEE Transactions on Multimedia*, **25**, 1674-1685. https://doi.org/10.1109/tmm.2022.3147369

[16] Zhou, H., Yu, J. and Yang, W. (2023) Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video

Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 3769-3777. https://doi.org/10.1609/aaai.v37i3.25489

[17] Lv, H., Yue, Z., Sun, Q., Luo, B., Cui, Z. and Zhang, H. (2023) Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 8022-8031. https://doi.org/10.1109/cvpr52729.2023.00775

[18] Xu, C., Xu, K., Jiang, X. and Sun, T. (2025) PLOVAD: Prompting Vision-Language Models for Open Vocabulary Video Anomaly Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, **35**, 5925-5938. https://doi.org/10.1109/tcsvt.2025.3528108